Aligning Modalities in Vision Large Language Models via Preference Fine-tuning

Year: 2024 | *Citations: 159* | *Authors: Yiyang Zhou, Chenhang Cui, Rafael Rafailov*

Abstract

Instruction-following Vision Large Language Models (VLLMs) have achieved significant progress recently on a variety of tasks. These approaches merge strong pre-trained vision models and large language models (LLMs). Since these components are trained separately, the learned representations need to be aligned with joint training on additional image-language pairs. This procedure is not perfect and can cause the model to hallucinate - provide answers that do not accurately reflect the image, even when the core LLM is highly factual and the vision backbone has sufficiently complete representations. In this work, we frame the hallucination problem as an alignment issue, tackle it with preference tuning. Specifically, we propose POVID to generate feedback data with AI models. We use ground-truth instructions as the preferred response and a two-stage approach to generate dispreferred data. First, we prompt GPT-4V to inject plausible hallucinations into the correct answer. Second, we distort the image to trigger the inherent hallucination behavior of the VLLM. This is an automated approach, which does not rely on human data generation or require a perfect expert, which makes it easily scalable. Finally, both of these generation strategies are integrated into an RLHF pipeline via Direct Preference Optimization. In experiments across broad benchmarks, we show that we can not only reduce hallucinations, but improve model performance across standard benchmarks, outperforming prior approaches. Our data and code are available at https://github.com/YiyangZhou/POVID.